BASIC NLP CONCEPTS

Legal text preprocessing transforms raw, unstructured legal documents into clean, structured formats suitable for NLP tasks like classification, entity extraction, and summarization. This is crucial in legal applications.

1) Sentence Segmentation

Imagine chopping a long legal document into bite-sized sentences so computers don't get lost in endless run-on paragraphs full of "wherefores" and citations.

2) Tokenization

This is like breaking a sentence into Lego bricks—words, phrases, or even citations as single pieces—instead of smashing them apart randomly. This step improves efficiency in classification and categorizing text.

3) Stop-Words Removal

It means to remove common words like 'the', 'is' while customizing lists for legal terms (e.g., excluding "whereas" in preambles), it reduces noise by 30-50% is vast datasets.

4) Lemmatization

Words in legal writing love changing forms—"judging," "judged," "judgments" all become plain "judge"—using smart dictionaries. It keeps meaning intact for better search and analysis in eDiscovery, avoiding goofy shortcuts that confuse AI.

5) Part-of-Speech Tagging

POS tagging assigns grammatical categories (e.g., noun, verb, adjective) to each token in legal texts, aiding dependency parsing and feature extraction for tasks like clause identification.

6) Named Entity Recognition

Spotting VIPs in the text—names like "John Doe," laws like "42 U.S.C. § 1983," or dates—turns chaos into a neat list for linking cases or flagging risks. After earlier steps, legal NER models nail 90%+ accuracy.